Regularization

We want to minimize empirical risk, but taken to an extreme we will overfit. So, we need some strategy to enforce parsimony and generality. We add an extra term that penalizes complexity. General form:

$$\min_{f \in \mathcal{F}} R_{\text{emp}}(f) + \lambda \Omega(f)$$

with λ the regularization parameter and Ω the regularizer.

Regularization for least squares

OLS tends to become overfit when D is large relative to n (and is completely useless if $D > n$). Ridge and lasso are techniques to regularize. SVD is useful for them.

Singular value decomposition (SVD)

Any $n \times p$ matrix $\mathbf{A}$ can be decomposed into the following:

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}'$$

Where $\mathbf{U}$ is a $n \times p$ matrix, $\mathbf{V}$ is a $p \times p$ matrix, and $\mathbf{D}$ is a $p \times p$ diagonal matrix.

Conceptually, SVD is similar to the eigen-decomposition $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}'$, where $\mathbf{A}$'s eigenvalues are the diagonal of $\mathbf{\Lambda}$ and $\mathbf{Q}$ is orthogonal. But the eigen-decomposition only is valid for positive definite matrices, whereas SVD is applicable for any matrix.

The diagonal values of $\mathbf{D}$ are the singular values of $\mathbf{A}$. They are the square root of the eigenvalues of $\mathbf{A}'\mathbf{A}$. All $d_1, d_2 ... d_p \geq 0$. If one ore more equals 0, $\mathbf{A}$ is singular and not of full rank. The ratio of the largest to the smallest d is called the condition number. A high condition number means an ill-conditioned matrix (i.e. one that is not numerically stable and may have results "blow up" from minor changes, perhaps from rounding off).

SVD and OLS

Let $\mathbf{U}_i$ be the i'th column of U, and similar for $\mathbf{V}$. Subscript r means the effective rank.

$$\hat{\beta} = \mathbf{V}_r \mathbf{D}^{-1} \mathbf{U}_r' \mathbf{y}$$

$\mathbf{V}_r \mathbf{D}^{-1} \mathbf{U}_r'$ is the pseudo-inverse of $\mathbf{A}$. This method is more numerically stable than the normal equations.

Also: the hat matrix can be computed via SVD.

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{U}\mathbf{U}'$$

Ridge regression (and SVD)

Stepwise and similar schemes are discrete, all-or-nothing. Ridge regression instead is continuous and shrinks the coefficients via a penalty on their size. Also known as weight decay in neural networks.

$$\hat{\beta}_{\text{ridge}} = \text{argmin}_\beta \left[\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}) + \lambda \sum_{j=1}^{n} \beta_j^2) \right]$$

$$\hat{\beta}_{\text{ridge}} = (y - X\beta)'(y - X\beta) + \lambda\beta'\beta = (X'X + I\lambda)^{-1} X'y$$

The fit (equivalent to $X\beta$ in OLS) can be expressed via SVD as:

$$X\beta_{\text{ridge}} = \mathbf{U}\text{diag}\left(\frac{d_1}{d_1 + \lambda} ... \frac{d_p}{d_p + \lambda}) \right) \mathbf{U}'\mathbf{y}$$

You have to normalize x_{ij} as $x_{ij} - \bar{x}_j$ and estimate $\beta_0 = \bar{y}$ (ridge doesn't estimate intercept).

Ridge makes the problem non-singular even if $\mathbf{X}$ is not of full rank.

Ridge works well if the first few eigenvalues are dominant and the rest small, and the regularization penalty chosen at the knee in the curve.

Ridge does not work well for sparse set-ups.

Lasso

Basic idea: ridge but with L1 norm. Zero norm (i.e. return 1 if $w_j \neq 0$) would be even better but is computationally intractable.

$$\hat{\beta}_{\text{lasso}} = \underset{\beta \in \mathbb{R}^d}{\text{argmin}} \frac{1}{2n} ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda ||\beta||_1$$

Naturally produces sparse solutions, i.e. is implicitly performing feature selection